SCALABLE PARALLEL APPROXIMATE FORMULATIONS OF MULTI-DIMENSIONAL SPATIAL AUTO-REGRESSION MODELS FOR SPATIAL DATA MINING[1]

Shashi Shekhar, Baris M. Kazar, David J. Lilja
Army High Performance Research Center
PO Box 581459
Minneapolis, MN 55458
Computer Science and Engineering Department
University of Minnesota
200 Union St. SE Minneapolis,
MN 55455 USA
{shekhar, kazar, lilja}@cs.umn.edu

ABSTRACT

The spatial auto-regression (SAR) model is a popular spatial data analysis technique which has been used in many applications with geo-spatial datasets. However, exact solutions for estimating SAR parameters are computationally expensive due to the need to compute all the eigen-values of a very large matrix. Therefore, serial solutions for the SAR model do not scale up to map sizes of interest to the Army. Thus, we developed the parallel approximate SAR models which can now be used by the Army to increase the accuracy and usefulness of maps, better analyze the impact of weather on the battlefield, make near-future predictions of the locations of enemy units, and increase the lethality of missiles.

1. INTRODUCTION

The Army generates, accesses, and manages large amounts of spatial data, resulting in the need for fast techniques to mine interesting patterns embedded in these very large size data. Linear regression, one of the best-known classical data mining techniques, cannot be applied to geo-spatial data, because its assumption of *independent identical distribution* (*IID*) in learning data samples does not hold true for geo-spatial data. In the SAR model [Shekhar et. al., 2003], spatial dependencies within data are taken care of by the auto-correlation term, and the linear regression model thus becomes a spatial auto-regression (SAR) model. Incorporating the auto-correlation term enables better prediction accuracy. However, computational complexity increases due to the logarithm of the determinant of a large matrix, which is computed by finding all of the eigen-values of another matrix. Thus, we developed parallel approximate SAR model solutions to reach the size of very large datasets.

Recently, the Topographic Engineering Center initiated a project that uses spatial data mining and the SAR model solution to help classify soil types for the Army. The approximate SAR models that we developed can also now be used by the Army with very large datasets in four other new areas. First, in recognition of the fact that maps are as important to soldiers as weapons, the SAR model solution can be used to identify errors in existing map attributes, detect unexpected correlations related to terrain properties, and as a prediction tool to fill in gaps in maps. Second, SAR model solutions can be used in battlefield weather impact applications including: a) learning rules to map, i.e., for correlations between humidity and fog, b) identifying boundaries between weather (systems).

In other examples, as a location prediction technique, parallel implementation of SAR can be used to predict near-future locations (global hot spots) of enemy units given current location based on a sensor network, battlefield terrain, historic war tactics, etc. Parallel implementation of SAR can also be used to increase the lethality of missiles via precision targeting using map-matching to evaluate flight trajectories for drifts from flight plan or need for revision due to unexpected obstacles.

Our contributions can be summarized as follows:

- Our parallel solution covers not only single but also multiple-dimensional problems i.e. 2-D, and 3-D geo-spaces for the SAR model solution.
- We offer portable software that can be run on multiple hardware platforms. We do not use any machine specific compiler directives in order to preserve portability.
- Ours is the first attempt to evaluate the scalability of SAR both analytically and experimentally.

[1] This work was partially supported by the Army High Performance Computing Research Center (AHPCRC) under the auspices of the Department of the Army, Army Research Laboratory (ARL) under contract number DAAD19-01-2-0014. This work received additional support from the University of Minnesota Digital Technology Center and the Minnesota Supercomputing Institute.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
00 DEC 2004	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Scalable Parallel Approximate Formulations Of Multidimensional Spatial Auto-Regression Models For Spatial Data Mining	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Army High Performance Research Center PO Box 581459 Minneapolis, MN 55458; Computer Science and Engineering Department University of Minnesota 200 Union St. SE Minneapolis, MN 55455 USA**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001736, Proceedings for the Army Science Conference (24th) Held on 29 November - 2 December 2005 in Orlando, Florida.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**2**	

2. PARALLEL APPROXIMATE SAR MODELS

Serial solutions for the exact SAR model do not scale up to map sizes of interest to the Army. For instance, it takes approximately one hour to solve a 10K problem size. Our research objective is to develop highly scalable approximate semi-sparse multi-dimensional parallel formulation of spatial auto-regression (SAR) models for location prediction problems using hybrid programming, and sparse matrix algebra in order to reach very large problem sizes i.e. billions of (observations). Hybrid programming enables greater scalability by using MPI across nodes and OpenMP within a single node. We developed an exact parallel spatial auto-regression model solution which is computationally dominated by calculations of the eigen-values of a large dense matrix and computes the maximum likelihood function. The exact solutions can run both sequentially and in parallel for medium-scale location prediction problems. We also developed a much faster and very accurate approximate parallel spatial auto-regression model solution based on Chebyshev polynomial approximation. This approximate spatial auto-regression model can exploit the sparse nature of the neighborhood matrix to save both execution time and memory.

3. RELATED WORK

To the best of our knowledge, there is only one other parallel implementation of the (exact) spatial auto-regression model solution, which is in one-dimension. The approach tries to solve by finding all of the eigen-values of a dense symmetric neighborhood matrix i.e. **W** for regular square tessellation one-dimensional planar surface partitioning. However, this parallel formulation used parallelized versions of the eigen-value subroutines from CMSSL, a parallel linear algebra library written in CM-Fortran (CMF) for the CM-5 supercomputers of Thinking Machines Corporation, neither of which is available for use anymore. Thus, it can be stated that the parallel formulation presented in this study is the only parallel spatial auto-regression formulation available in the literature. Furthermore, the spatial auto-regression model solution presented here is more generic, harder to solve, and covers not only single but also multi-dimensional geo-spaces as well.

4. EXPERIMETAL RESULTS

Figure 1 shows the serial and parallel execution times (in seconds) of computing the logarithm of the determinant of a matrix in the SAR model solution using Chebyshev approximation. Computing all of the eigen-values of the neighborhood matrix, which is used to compute the logarithm of the determinant of a matrix, takes approximately 99% of the total serial response time [Kazar et al., 2003,2004,2004] and up to an hour to compute a 10K problem size. By contrast, as shown in the figure, it takes only slightly more than 16 seconds to solve the same problem size using Chebyshev polynomial approximation. Chebyshev approximation thus not only scales much better than the eigen-value based approach in terms of execution time but also uses very little memory [Kazar et al. Oct.2004].

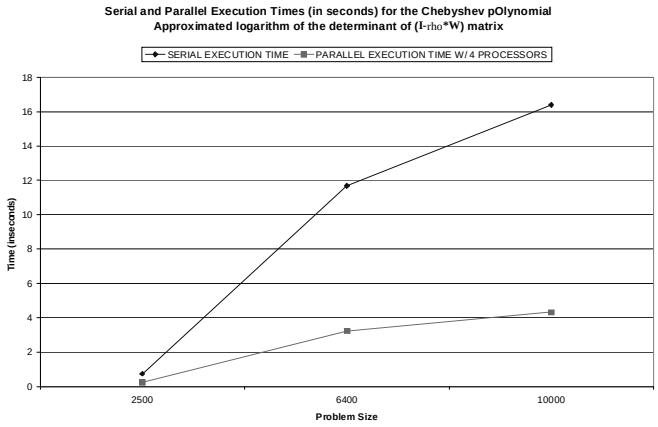

Fig. 1. The serial and parallel execution times (in seconds) for the Chebyshev polynomial approximated logarithm of the determinant of matrix $(\mathbf{I} - \rho\mathbf{W})$. The problem size is varied as 2500, 6400 and 10000. The number of processors is kept at four for parallel execution times.

5. CONCLUSIONS

Incorporating the auto-correlation term in the SAR model enables better prediction accuracy with respect to linear regression. However, computational complexity increases due to the logarithm of the determinant of a large matrix. In order to reach very large dataset sizes the Army uses, we developed scalable approximate SAR model solution in this study.

REFERENCES

Shekhar, S., Chawla, S., 2003: Spatial Databases: A Tour, Prentice Hall 2003.

Kazar, B., Shekhar, S., Lilja, D., 2003: Parallel Formulation of Spatial Auto-regression, AHPCRC Technical Report No: 2003-125 (Poster at PPoPP 2003)

Kazar, B. M., Shekhar, S., Lilja, D. J., Boley, D., 2004: A Parallel Formulation of the Spatial Auto-Regression Model for Mining Large Geo-Spatial Datasets, AHPCRC Technical Report no. 2004-103 (SIAM HPDM Workshop 2004)

Kazar, B. M., Shekhar, S., Lilja, D. J., Vatsavai, R. R., Pace, R. K., 2004: Comparing Exact and Approximate Spatial Auto-Regression Model Solutions for Spatial Data Analysis, Third International Conf. on Geographic Information Science (GIScience2004), October 2004, Maryland,USA